AI 赋能的网络攻击分析与分类

Zhi WANG
University of Chinese Academy of Sciences

摘要

近些年，伴随着计算机性能的不断提升，人工智能（Artificial Intelligence，AI）技术得以迅速发展。以深度神经网络为代表的人工智能技术在包括自动驾驶、智慧城市、医学图像等在内的多个领域取得了巨大突破，并开始走进千家万户。在 AI 惠及大众的同时，我们也注意到 AI 能够助力攻击者执行网络攻击任务。本文对已有的 AI 赋能网络攻击案例进行梳理，分析 AI 在攻击任务中的作用，将 AI 赋能的网络攻击划分为在线和离线两类，并给出相应的讨论与展望。

关键词 人工智能 网络攻击 神经网络 网络空间安全 人工智能安全

1 引言

人工智能作为引领未来的战略性技术，日益成为驱动经济社会各领域从数字化、网络化向智能化加速跃升的重要引擎。近年来，数据量爆发式增长、计算能力显著性提升、深度学习算法突破性应用，极大地推动了人工智能发展。然而，技术往往具备两面性。当技术应用于好的一面时，能推动科技发展和社会进步；而当其应用在不好的一面时，能造成发展停滞和社会动荡。人工智能技术也是如此。当应用在网络安全领域时，人工智能可以在入侵检测、恶意代码防御、态势感知等多个方向提供有力支撑，也可以助力攻击者执行更加高效和难以防御的网络攻击。

2018 年，由来自 26 所不同研究机构的科学家联名发布了针对恶意利用人工智能的报告，提出了恶意利用 AI 可能导致的数字安全、物理安全和政治安全风险，号召全世界的 AI 研究团队警惕 AI 安全风险，抵制针对 AI 的恶意利用行为。随着 AI 安全风险的加剧，一些知名安全厂商也对 AI 应用于网络攻击的趋势进行了预测。BeyondTrust[1]在网络安全发展趋势预测中指出，机器学习训练数据污染、AI 武器化的泛滥将给网络安全带来巨大挑战。Check Point[2]在网络威胁趋势预测中认为，AI 技术呈现武器化趋势。Gartner[3]则表示，到 2022 年，30%的网络攻击将与人工智能安全有关，人工智能趋于工程化。Fortinet[4]也认为，借助于多形态恶意代码的进化、集群攻击和人工智能的武器化成为趋势。尽管人工智能可以运用在多个领域（如自动驾驶、武装器械、智能家居等），本文主要关注其在网络安全领域的应用，侧重点在网络层及以上，聚焦于应用层。

近些年，我们见到了一些 AI 应用于网络攻击方面的案例。为了更好地研究 AI 赋能的网络攻击所具有的特性，更有效地进行针对性的防御工作，有必要对 AI 赋能的网络攻击案例进行梳理，有必要研究 AI 在网络攻击中起到的作用。为此，本文将梳理相关案例，分析攻击类别，研究 AI 作用，提出分类方法，并对 AI 赋能的网络攻击发展趋势进行展望。

2 攻击案例回顾

本章对部分有代表性的 AI 赋能网络攻击的案例进行回顾。

2016-2019

2016 年，Seymour 等[5]在全球著名的黑客大会 BlackHat 上提出借助人工智能技术在推特上进行高度定制化的自动化鱼叉式钓鱼的方法，该方法使用聚类对高价值的目标进行筛选，结合自然语言处理领域的方法分析目标感兴趣的话题，构建 SNAP_R（社交网络自动侦察钓鱼，Social Network Automated Phishing with Reconnaissance）基于 LSTM 生成鱼叉式钓鱼内容，并能根据目标活动时间推送给目标，引诱目标上钩。

Sivakorn 等[6]在 BlackHat 上提出了自动化绕过谷歌 reCAPTCHA 的方法。借助对图片的描述词，对给出的图片进行切割并利用在线平台进行检索，通过使用 NLP 方法提取页面返回的信息计算和描述词的相似度，可判断所选图片是否为目标图片。当外部网络不可达时，可以使用离线模型进行判断。

Anderson 等[7]提出 DeepDGA，提出结合自编码器（Auto-Encoder）和长短期记忆网络（Long Short-Term Memory）学习 Alexa Top 100M 域名的特征，将编码器和解码器在生成对抗网络（检测器+生成器）中重新组装，生成仿真度极高的 DGA 域名，能够绕过 DGA 检测器。

2017 年，Baki 等[8]利用希拉里“邮件门”事件中泄露的电子邮件，使用自然语言处理的技术分析邮件文法特征，生成（伪造）一批来自希拉里和佩林的邮件让其他人进行辨别。结果显示，多数人认为生成的邮件来自于希拉里和佩林，其中因语言通俗等原因，来自“希拉里”的邮件得到了更多的票数。该项技术可用于鱼叉式钓鱼，生成高仿真的钓鱼邮件。

Hu 等[9]提出使用生成对抗网络算法产生对抗恶意软件样本，使用生成模型给恶意样本添加扰动，用替代检测器判断样本是否为恶意，替代检测器由良性样本和生成模型生成的对抗样本训练而来，起到模拟黑盒检测器的作用。通过对抗训练的过程，能够最小化生成对抗样本被检出的概率。

2018 年，Kirat 等[10]在 BlackHat 上提出了 DeepLocker，用于定向隐蔽攻击。DeepLocker 根据攻击目标信息训练神经网络模型，用模型的输出结果将作为对称密码算法的密钥，加密恶意载荷。当发现目标时，神经网络模型能够输出正确解密密钥解开载荷执行恶意

功能。当目标未出现时，由于神经网络模型的单向性和抗碰撞性，分析人员无法构建触发条件获取密钥，进而抵抗分析。

Bahnsen 等[11]在 BlackHat 上提出自动化钓鱼工具 DeepPhish，使用 LSTM 学习 URL 的特征，并构造恶意 TLS 证书构建钓鱼页面。与普通方法对比，DeepPhish 能使钓鱼成功率提升 20%到 30%。

Takaesu[12]提出自动化渗透测试攻击 DeepExploit，使用强化学习对给定的目标进行扫描和探测，并根据探测结果利用 Metasploit 进行自动化的渗透测试，在测试结束后生成测试报告。

Anderson 等[13]提出利用强化学习的激励和反馈机制修改静态 PE 恶意代码的结构特征来规避检测的方法，使用黑盒检测器作为样本修改结果的反馈源，通过增加无用函数调用、修改段名、移除签名信息等方式得到能逃逸检测的样本。

Ye 等[14]提出一种新的文本类验证码破解方法。与已有使用大量手动标记的真实验证码进行学习的破解器不同，这里采用生成对抗网络的方式，用已有的数据库训练验证码模型，当遇到新的验证码时，使用迁移学习方法，仅需要少量的新验证码数据就能得到很好的破解效果。

Rigaki 等[15]提出使用生成对抗网络模拟 Facebook 聊天的流量来逃避基于流量分析的检测器，该方法使用 Stratosphere Linux IPS 做检测器来检测流量是否为 Facebook 聊天流量，使用 RNN 和 LSTM 构建生成器和判别器，只有当检测器认为所生成的流量为 Facebook 聊天流量时才予以放行，用训练出的参数在恶意代码活动中进行流量生成来绕过基于流量的检测器。

2020-2021

2020 年，Li 等[16]提出在特征空间中生成对抗性特征向量，将向量转换为对抗性恶意 PDF 组件，修改 PDF 文件 CRT 与尾部之间的特征，能生成逃避 PDF 恶意代码检测器的恶意 PDF 文件。通过在 PDFRate 分类器上评估该方法，能在四种逃避情况下对目标系统知识进行攻击。

Novo 等[17]提出了一个基于流的 C2 流量检测器和基于代理的 C2 流量规避检测方法，使用快速梯度符号法（Fast Gradient Sign Method，FGSM）用白盒方法在保证数据通信功能的条件下，打破 C2 数据流的统计特征，对抗恶意 C2 流量检测器。

Wang 等[18]提出使用神经网络进行 C2 寻址的方法 DeepC2，借助神经网络模型的不可逆性，使用神经网络模型识别控制端账号的头像进行寻址，同时使用数据增强技术解决以往基于社交网络平台的 C2 中的异常内容问题。

Yuan 等[19]提出了端对端字节级黑盒对抗攻击方法，用生成器生成一段 Payload 添加在样本后，通过黑盒检测器的反馈结果训练鉴别器。在使用时，只需要生成器生成 Payload 即可对抗检测器。此方法能在添加的 Payload 长度为 2.5%时达到 100%的逃逸成功率。

XunSu 等[20]提出使用自然语言处理和 LSTM 批量自动化探测 WAF 规则的方法，通过数据加算法加探测的方式，自动化提取 Payload 的文本特征，采用注意力机制增强关键词学习，根据经验自动化方法确定在不同位置上使用的不同关键字，能线上反馈探测打击边界，实现 WAF 绕过。

2021 年，Wang 等[21]提出在神经网络模型中嵌入恶意代码来进行攻击载荷投递的方法 EvilModel，通过分析神经网络模型的结构，将神经网络的神经元及参数替换为恶意代码，可以在不影响模型性能的条件下，将大体积的恶意代码免杀地投递到目标设备上。

3 攻击分类

上一章对部分有代表性的 AI 赋能网络攻击的案例进行了回顾。本章针对 AI 在这些网络攻击中起到的作用进行分析。

根据攻击场景的不同，上述案例可以分为以下几个方面：

- 自动化生成。典型的案例有推特及邮件钓鱼[5][8]、恶意样本生成[9]、恶意流量生成[15][17]等。
- 自动化攻击。典型的案例有 DeepExploit[12]等。
- 欺骗。典型的案例有各种钓鱼攻击，如推特钓鱼[5]、邮件钓鱼[8]、DeepPhish[11]等。
- 检测逃逸。典型的案例有恶意域名检测器的逃逸[7]、恶意软件检测器的逃逸[13][19][21]、恶意流量检测器的逃逸[15]等。
- 破解。典型的案例有人机交互图灵测试的破解[6][14]、防火墙规则的破解[20]等。
- 目标识别。典型的案例有识别攻击对象[10]、识别攻击者[18]等。

在不同的场景中，AI 发挥着不同的作用。Minsky 等[22]结合 AI 具备的能力，将 AI 在网络攻击中的作用总结如下：

- 预测，即根据已有数据进行预测。
- 生成，即根据目标情况生成新的内容。
- 分析，即从已有数据中挖掘有效信息。
- 检索，即从已有数据中搜索报告目标信息。
- 决策，即根据已有信息对下一步行动给出指导。

根据对 AI 的使用不同，可以将 AI 赋能的网络攻击分为“在线”和“离线”两类：

- “在线”（online）指 AI 模型直接用在攻击过程中，其输出结果直接用于在执行中的攻击任务。
- “离线”（offline）指 AI 模型处于攻击活动的后方，其输出结果可用于攻击任务中，也可不用于攻击任务中。

具体地，“在线”聚焦于 AI 模型和方法本身的特性，利用 AI 与其他方法不同的特点，将其应用于网络攻击的特定任务中；“离线”聚焦于 AI 能干什么、能用在网络攻击的哪个环节，其主要任务是将人能做到的事情自动化、智能化，比人工要快要好。“在线”模式中，AI 模型处于目标环境中，而“离线”模式中，AI 模型处于攻击者的环境中。按照“在线”和“离线”的定义和划分，现有的大部分案例都属于“离线”范畴，如下：

- “在线”模式典型案例：DeepLocker、DeepC2、EvilModel、集群智能等。
- “离线”模式典型案例：恶意内容生成类、欺骗类、破解类、自动化攻击类等。

无论是“在线”还是“离线”，上述案例都是 AI 赋能的网络攻击的重要组成部分，如表 1。

表 1 AI 赋能的网络攻击模式分类

攻击模式	内容	AI 作用	典型案例
在线	AI 模型直接用在攻击过程中，其输出结果直接用于在执行中的攻击任务。	生成、检索、决策	DeepLocker、DeepC2、EvilModel、集群智能等。
离线	AI 模型处于攻击活动的后方，其输出结果可用于攻击任务中，也可不用于攻击任务中。	预测、生成、分析、决策	恶意内容生成类、欺骗类、破解类、自动化攻击类等。

4 讨论与展望

从严格意义上说，AI 赋能的网络攻击不属于“人工智能安全”范畴，而是“人工智能+安全”范畴，属于人工智能在网络安全领域（尤其是网络攻击领域）的应用。纯粹的“人工智能安全”是指人工智能自身的安全问题，涉及到人工智能模型安全、人工智能数据安全、人工智能框架安全等，腾讯的相关团队[23]对已有工作做了总结，如表 2。此外，AI 应用于网络安全的检测与防护也属于人工智能的应用，同属于“人工智能+安全”。将攻防两个子方向合并，可得到人工智能安全的更完整的划分，如图。

提及人工智能自身的安全问题，就要引申出新的一类攻击，即针对人工智能的弱点，对部署有人工智能装置的智能化系统的攻击。比如已有攻击案例中的检测逃逸类，其目标就是部署有基于人工智能方法的恶意目标检测器。攻击者也可以对已有的智能化系统进行模型窃取、数据投毒、对抗样本等攻击，干扰系统的正常运行。此类攻击正处于发展阶段。可以预见的，未来几年，相关领域的成果将不断出现。

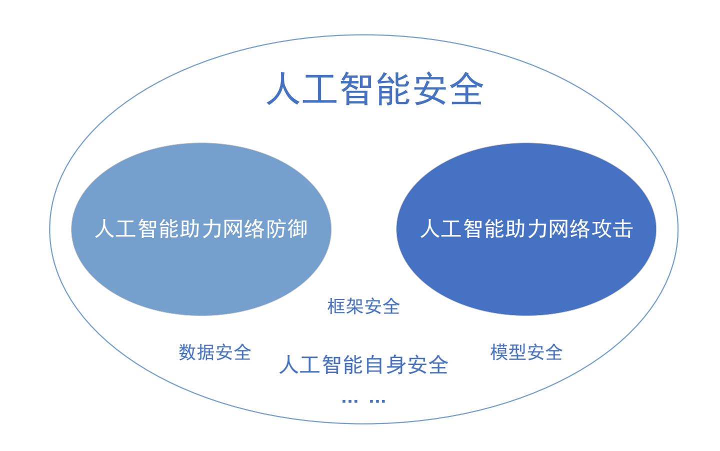


图 人工智能安全体系

在应用上，囿于 AI 模型运行所需要的各种资源（算力、数据、模型等），AI 赋能的网络攻击（尤其是“在线”模式的网络攻击）尚且不能应用于真实世界的网络攻击任务中。随着 AI 的进一步推广和普及，未来计算设备可能集成相关的资源，使攻击者得以实施此类攻击。因此，相关领域的研究人员有必要提高警惕，预先对此类攻击进行有针对性的防护工作，进一步保护和提升各类信息系统的安全性，在此类攻击出现的第一时间内做到有效防御。

表 2 人工智能自身安全风险[23]

环境接触	数据搜集整理	模型训练	模型部署	模型使用	模型架构	结果影响
依赖软件攻击	数据投毒	梯度中恢复数据	模型中数据恢复	数字对抗攻击	查询式架构窃取	模型误判
Docker 恶意访问	数据后门攻击	初始权重修改	模型文件攻击	物理对抗攻击	侧信道架构窃取	信息泄露
硬件模型后门攻击		代码攻击		模型窃取		
供应链攻击		训练后门攻击		GPU/CPU 溢出破坏		
		非集中式部署				

5 总结

本文针对 AI 赋能的网络攻击进行了梳理和分类。本文首先分析了网络安全形势，筛选了近些年具有代表性的 AI 赋能的网络攻击案例，对相关工作进行了简要的介绍。随后，本文研究和分析了 AI 在相关案例中的作用，并根据各类攻击任务对 AI 的使用不同将 AI 赋能的网络攻击分为“在线”和“离线”两种模式，对两种模式进行介绍。最后，本文对人工智能和网络安全的结合应用做了讨论和展望。网络攻击与防御是相生相伴的技

术，二者此消彼长，相互促进。我们有理由认为，未来的信息系统内会集成针对 AI 赋能的网络攻击的防护，未来的防御手段也会借助于智能化的力量更加强大、更有力地保障网络空间安全。

致谢

在完成本文的过程中，ArkTeam 和山城安全给予了很大帮助，在此致谢。

参考文献

1. BEYONDTRUST. Beyondtrust releases cybersecurity predictions for 2021 and beyond[EB/OL]. 2020. https://www.globenewswire.com/news-release/2020/10/28/2115996/0/en/BeyondTrust-Releases-Cybersecurity-Predictions-for-2021-and-Beyond.html.
2. CHECKPOINT. Check Point Software´s predictions for 2021: Securing the ‘next normal’[EB/OL]. 2020. https://blog.checkpoint.com/2020/11/10/check-point-softwares-predictions-for-2021-securing-the-next-normal/.
3. CEARLEY D, JONES N, SMITH D, et al. Top 10 Strategic Technology Trends for 2020[EB/OL]. 2019. https://emtemp.gcom.cloud/ngw/globalassets/en/doc/documents/432920-top-10-strategic-technology-trends-for-2020.pdf.
4. FORTIGUARD. New cybersecurity threat predictions for 2021[EB/OL]. 2020. https://www.fortinet.com/blog/threat-research/new-cybersecurity-threat-predictions-for-2021.
5. SEYMOUR J, TULLY P. Weaponizing data science for social engineering: Automated e2e spear phishing on twitter[J]. Black Hat USA, 2016, 37:1-39.
6. SIVAKORN S, POLAKIS J, KEROMYTIS A D. I’m not a human: Breaking the Google reCAPTCHA[J]. Black Hat, 2016:1-12.
7. ANDERSON H S, WOODBRIDGE J, FILAR B. Deepdga: Adversarially-tuned domain generation and detection[C]//Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2016, Vienna, Austria, October 28, 2016. ACM, 2016: 13-21.
8. BAKI S, VERMA R M, MUKHERJEE A, et al. Scaling and effectiveness of email masquerade attacks: Exploiting natural language generation[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017. ACM, 2017: 469-482.
9. HU W, TAN Y. Generating adversarial malware examples for black-box attacks based on GAN[J/OL]. CoRR, 2017, abs/1702.05983. http://arxiv.org/abs/1702.05983.
10. KIRAT D, JANG J, STOECKLIN M. Deeplocker–concealing targeted attacks with ai locksmithing[J]. Blackhat USA, 2018.

11. BAHNSEN A C, TORROLEDO I, CAMACHO L D, et al. Deepphish: Simulating malicious ai[C]//2018 APWG Symposium on Electronic Crime Research (eCrime). 2018: 1-8.
12. TAKAESU I. Deep exploit: Fully automatic penetration test tool using machine learning[J]. Blackhat EUROPE, 2018.
13. ANDERSON H S, KHARKAR A, FILAR B, et al. Learning to evade static PE machine learning malware models via reinforcement learning[J/OL]. CoRR, 2018, abs/1801.08917. http://arxiv.org/abs/1801.08917.
14. YE G, TANG Z, FANG D, et al. Yet another text captcha solver: A generative adversarial network based approach[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018. ACM, 2018: 332-348.
15. RIGAKI M, GARCIA S. Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection[C]//2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018. IEEE Computer Society, 2018: 70-75.
16. LI Y, WANG Y, WANG Y, et al. A feature-vector generative adversarial network for evading pdf malware classifiers[J]. Information Sciences, 2020, 523: 38 - 48.
17. NOVO C, MORLA R. Flow-based detection and proxy-based evasion of encrypted malware c2 traffic[C]//AISec'20: Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security. New York, NY, USA: Association for Computing Machinery, 2020: 83–91.
18. WANG Z, LIU C, CUI X, et al. DeepC2: AI-powered convert botnet command and control on OSNs[J/OL]. CoRR, 2020, abs/2009.07707. http://arxiv.org/abs/2009.07707.
19. YUAN J, ZHOU S, LIN L, et al. Black-box adversarial attacks against deep learning based malware binaries detection with GAN[C]//Frontiers in Artificial Intelligence and Applications: volume 325 ECAI 2020 - 24th European Conference on Artificial Intelligence, Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020). IOS Press, 2020: 2536-2542.
20. XUNSU, KEYUNLUO. Deep X-Ray: 一种机器学习驱动的 WAF 规则窃取器[EB/OL]. 2020. http://t.cn/A65ZGOyL.
21. Wang Z, Chaoge Liu, and Xiang Cui. EvilModel: Hiding Malware Inside of Neural Network Models[C]. In 2021 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2021: 1-7.
22. MIRSKY Y, DEMONTIS A, KOTAK J, et al. The Threat of Offensive AI to Organizations[J/OL]. CoRR, 2021, abs/2106.15764. http://arxiv.org/abs/2106.15764.
23. 腾讯 AI Lab, 腾讯安全平台部朱雀实验室. AI 安全的威胁风险矩阵[OL]. 2020. https://matrix.tencent.com/